Detecting Short Adjacent Repeats in Multiple Sequences: A Bayesian Approach

LI, Qiwei

A Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Master of Philosophy

in

Information Engineering

The Chinese University of Hong Kong

August 2010

Abstract of thesis entitled:

Detecting Short Adjacent Repeats in Multiple Sequences: A Bayesian Approach

Submitted by LI, Qiwei

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in August 2010

The detection of repeating patterns in sequences is of interest in many fields. Biological sequences are especially enriched of many types of repeating patterns. For example, repetitive DNA sequences occur frequently in genomes. In this thesis, we study the detection of short adjacent repeats in multiple sequences. As a type of repetitive DNA sequences, a short adjacent repeat is an array of two or more approximate copies of a pattern, where gaps might exist between any neighboring repeating units. Identifying short adjacent repeats is an important issue in Bioinfor-

matics. Take tandem repeats, a special type of short adjacent repeats where the length of any inter-unit gap is zero, for example, researches have suggested that tandem repeats play a significant role in genetic markers, gene regulation, and human diseases.

Although the problem of tandem repeats detection has been widely studied, traditional methods, which are either based on string matching methods or signal processing methods, largely rely on the periodicity of a short segment. Therefore, they cannot handle gaps between repeating units. They are also incapable of detecting the common sequence pattern in multiple DNA sequences. This problem is becoming more serious as more genomes become available. One motivation of this research is to detect the common short adjacent repeats shared by multiple DNA sequences from some particular loci because it might shed light on molecular structure, function, and evolution.

In this thesis, we first formulate short adjacent repeats detection as a statistical problem by using a full probabilistic generative model. We then propose a Markov chain Monte Carlo

(MCMC) algorithm to infer the parameters of the model in a de novo fashion. The solution differs from all existing methods since it is of probabilistic nature and the statistical inference gives us a comprehensive picture of the related uncertainty.

As the features of short adjacent repeats are motif pattern, location, and structure, the key idea of our method is to use the motif matrix Θ to model repeating units as stochastic strings which are adjacently embedded in a homogeneous background Φ, and to implement a Bayesian approach to infer the location set $\mathbf{A}$ and the structure set $\mathbf{S}$ in the missing-data framework. We specify an independent prior distribution for each parameter. Then, the MCMC algorithm iterates the sampling steps, each of which updates the corresponding parameter conditional on the current values of all other parameters. After the MCMC chains converge, the sampled parameter values give us a whole picture of their jointly posterior distribution. We demonstrate the effectiveness of our scheme using both synthetic data and real biological data.

中文摘要

重複序列檢測在許多領域中都是一類令研究者們感興趣的問題。特別是在生物序列中，富含很多重複序列。例如，在基因組中有許多重複 DNA 序列。此篇論文主要研究在多條 DNA 鏈中檢測串聯重複序列的問題。作為重複 DNA 序列的一類，串聯重複序列是一組由兩個或以上相似的重複單元串聯而成的序列。有研究表明，串聯重複序列在遺傳標記，基因調控，與人類疾病中扮演著重要的角色。因此，在生物訊息學中，檢測串聯重複序列是一類非常重要的問題。

儘管該類問題已被廣泛研究，但無論是基於字符串匹配的方法，亦是基於信號處理的方法，都在很大程度上依賴於重複單元的週期性。因此，如果一些重複單元之間存在間隔，那麼這些方法都不能妥善處理之。而且，這些方法天生沒有能力解決在多條鏈中識別共有串聯重複序列的問題。尤其是現在，隨著越來越多生物基因組測序工作的完成，我們迫切需要一種能在多條鏈中檢測共有串聯重複序列的方法，因為這些共有的特定位點可能有助於幫助科學家闡明分子結構，功能及其進化。

在這篇論文中，我們將串聯重複序列的檢測轉化為一個統計學問題。首先，我們建立了一個完全基於概率的生成模型。然後，我們提出了一種馬爾可夫鏈蒙特卡羅（MCMC）演算法來推斷該模型中的參數。該解決方案不同于現有的方法是因為它與生俱來的概率特性。此外，統計推論使得我們對問題中的相關不確定性有一個較全面的認識。

串聯重複序列有三個特徵：重複單元，起始位置和內部結構。此方法的主要思想是利用主題矩陣 Θ 表示嵌入於均勻的背景 Φ 中的重複單元，並且在資料缺失的框架下利用貝葉斯（Bayesian）方法推斷表示起始位置的參數向量 **A** 與表示內部結構的參數矩陣 **S**。我們為每一個參數都設定先驗概率分佈。然後，MCMC 演算法依次反復迭代運算以便有條件地更新各個參數。當此 MCMC 鏈收斂後，記錄下來的所有抽樣值會展現給我們一個全部參數的聯合後驗概率分佈。我們分別使用合成數據和

真實生物數據證明此方法的有效性。

Publications

[1]. **Q. Li**, T. Liang, S. -Y. R. Li and X. Fan, "Bayesian Approach for Identifying Short Ajacent Repeats in Multiple DNA Sequences," The 2010 International Conference on Bioinformatics & Computational Biology (BIOCOMP'10), 2010, vol. 1, pp. 255-261;

[2]. **Q. Li**, T. Liang, S. -Y. R. Li and X. Fan, "Detection of Tandem Repeats in Multiple DNA Sequences via Probabilistic Approach," The 2010 ACM-HK Bioinformatics Symposium (HKUST), 2010, accepted;

[3]. Y. Li, M. Chen, **Q. Li**, and W. Zhang, "Enabling Multi-level Trust in Privacy Preserving Data Mining," IEEE Transaction on Knowledge and Data Engineering, 2010, in revision;

[4]. **Q. Li**, T. Liang, X. Fan, C. Xu, W. Yu, and S. -Y. R. Li, "An Automatic Procedure to Search Highly Repetitive Sequences in Genome as Fluorescence in Situ Hybridization Probes and Its Application on *Brachypodium Distachyon*," The 2010 IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010), 2010, submitted;

[5]. **Q. Li**, X. Fan, T. Liang, and S. -Y. R. Li , "Markov Chain Monte Carlo Algorithms for Detecting Short Adjacent Repeats in Multiple Sequences," Bioinformatics, 2010, to be submitted;

[6]. T. Liang, **Q. Li**, X. Fan, and S. -Y. R. Li , "Detection of Short

Dispersed Tandem Repeats by Reversible Jump Markov Chain Monte Carlo," Journal of Computational and Graphical Statistics, 2010, to be submitted;

[7]. J. Xu, **Q. Li**, T. Liang, V. O. K. Li, X. Fan, and S. -Y. R. Li, " An Evolutionary Monte Carlo Algorithm for Identifying Short Adjacent Repeats in Multiple Sequences," The 2010 IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010), 2010, to be submitted;

Acknowledgement

I am very fortunate to have the great scholar, Professor Shuo-Yen Robert Li as my supervisor. During the past two years, he has always supported me in every respect. I never forget his earnest teaching in my first postgraduate course in the university and his sportsmanlike attitude in my first hiking in Hong Kong. His enthusiasm for top quality research will surely inspire me through my life. I sincerely thank him for his great guidance and continuous support.

I would like to take this opportunity to express my sincere gratitude to my co-supervisor, Professor Xiaodan Fan. It is he that leads me into the wonderful land of Bioinformatics. His profound knowledge of both statistics and engineering has offered me lots of help through the whole research work. His

rigorous manner towards research has encouraged me to seek truth. I never forget he revised and polished my first paper over ten consecutive hours before the deadline coming. Without his contributions, I could have never completed this research work.

The Switching Lab has been a fantastic place for both research and living. I thank my present and past officemates, Dr. Jian Zhu, Dr. Xuesong Tan, Mr. Zhengfeng Qian, Dr. Qifu Sun, Dr. Siu-Ting Ho, Mr. Ziyu Shao, Mr. Zizhou Wang, Ms. Xiaoming Wu, Ms. Shuqin Li, Ms. Lok-Man Law, and Mr. Tong Liang, for their help and support.

I want to thank my father, my mother, and my younger brother for their unconditional support and trust. I devote the most special gratitude to my fiancee Ms. Na Kang, whose love makes my life a joyful journey. I might not have achieved this thesis without them.

I am also grateful to all of my relatives and friends. Thanks for their encouragement, concern, and help.

Last but not least, I would like to thank the Chinese University of Hong Kong and the Department of Information Engi-

neering.

This research is supported by the grants from the Research Grants Council of the Hong Kong S.A.R. (Grant No. CUHK 400709) and a CUHK direct grant (Grant No. CUHK 2060362).

This work is dedicated to my beloved fiancee Na.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Summary

In this thesis, we mainly focus on the algorithm that detects short adjacent repeats in multiple DNA sequences. In the first chapter, we answer the following questions: what is the repetitive DNA sequence, what are the tandem repeats, what are the short adjacent repeats, what is the significance of studying tandem repeats, and what are the main contributions of this thesis. In the end of this chapter, the layout of this thesis is given.

Since J. D. Watson and F. Crick discovered the double helix

structure of DNA molecules in 1953, the modern era of molecular and structural biology has come. A single-strand DNA sequence is a succession of four nucleotides, denoted by the four letters A, T, C and G, representing the primary structure of a real DNA molecule. It is this quaternary system that determines the protein functions and carries the inheritable genetic information which constitutes the genetic blueprint of living organisms. Over the past twenty years, the Genome Projects has generated abundant of genome sequence data (the whole set of DNA sequences of a species) whose size varies from about $5,000$ base pair (bp) in a very simple organism (e.g., the viruses SV40) to more than 10^{11} bp in some higher plants; the human's genome contains about 3×10^9 bp [19].

Nowadays, the availability of abundant of finished genome sequence data motivates the research works on DNA sequence analyses, including gene searching, restriction enzyme cutting site searching, transcription factor binding site searching, repeats detection, composition detection, etc. One of the popular research areas is repeats detection. In the past thirty years, the

identification of repetitive patterns in biological sequences has been an interesting, but challenging, problem. In this thesis, we mainly focus on the algorithm that detects short adjacent repeats in multiple DNA sequences.

1.1 Repetitive DNA Sequence

1.1.1 Definition and Categorization of Repetitive DNA Sequence

Repetitive DNA, which can be found in all living organisms, is a type of DNA sequences that are repeated many times in a haploid genome [45]. Noted that it is unnecessary that they are exactly the same, but they consist of families of sequences related [22]. In some organisms, repeats even make up a substantial fraction of the entire genome [5], e.g., 44.9% of the human genome is occupied by DNA repeats [18].

Repetitive DNA can be divided into two categories: tandem repeats and interspersed repeats [5]. The former ones are those repeats of which repeat units are placed next to each other in

an array, while the latter ones are those repeats of which repeat units are distributed around the genome in an apparently random fashion. Figure 1.1 shows the classification system of repetitive DNA sequence [5]. As shown in Figure 1.1, we mainly concentrate on the left branch of this repetitive DNA classification tree.

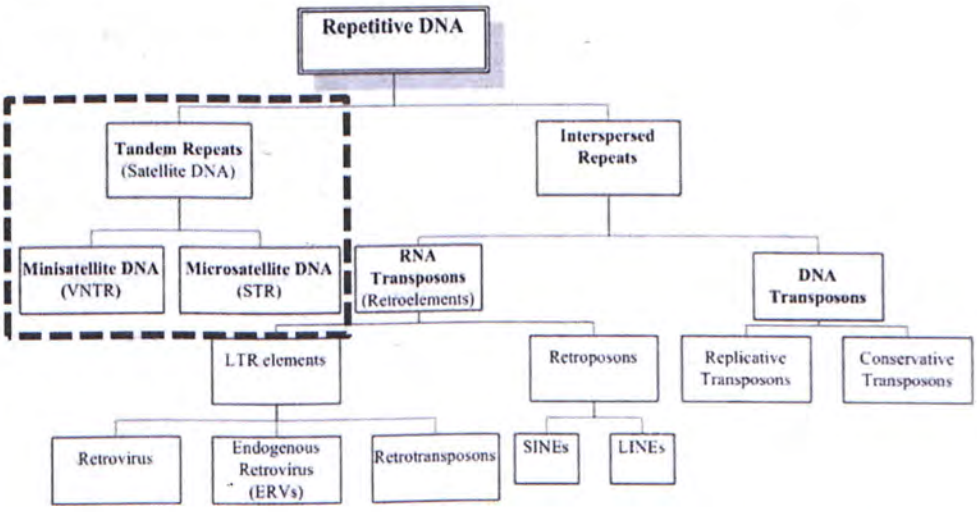

Figure 1.1: Repetitive DNA sequence classification system.

1.1.2 Definition and Categorization of Tandem Repeats

A tandem repeat in DNA is two or more contiguous, approximate copies of a pattern of nucleotides [2]. An example would be shown in Figure 1.2, where the pattern (also called the repeat

unit) $ATCCG$ is repeated three times.

Repeats location, repeats structure, pattern width, and copy number are four major features of tandem repeats [50]. The tandem repeats can be classified into two types according to the last two features [38]: minisatellites (also called variable number of tandem repeats (VNTR)) and microsatellites (also called simple tandem repeats (STR)). The size of minisatellites ranges from $1k$ bp to $20k$ bp, where the pattern width ranges from 9 bp to 80 bp. The microsatellites are less than 150 bp long, whose pattern width ranges from 2 bp to 6 bp. Table 1.1 briefly lists the tandem repeats classification system and their examples.

Since the widely used definition of tandem repeats emphasizes the consecutiveness, we erase this constraint by allowing short gaps between repeating units. Hence, we consider tandem repeats as a special type of short adjacent repeats where the length of any gap is zero.

Table 1.1: Tandem repeats classification system.

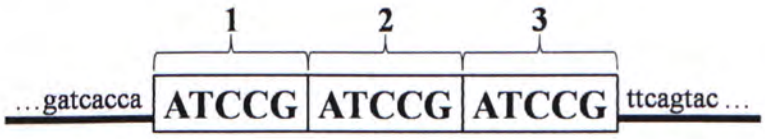

Figure 1.2: An example of a tandem repeat existing in a DNA sequence. Its pattern has a width of 5 bp and a copy number equal to 3. Lower case letters represent background nucleotides.

Type	Size / Pattern width	Example	Location
Minisatellites	$1k - 20k$ bp / $9 - 80$ bp	Hypervariable minisatellites	Centromeric regions
		Telomeric minisatellites	Near the telomeres
Microsatellites	≤ 150 bp / $1 - 6$ bp	$(CA)_n$	All chromosomes

1.1.3 Definition and Categorization of Interspersed Repeats

Different from tandem repeats, interspersed repeats are arisen by transposition, which is the movement of a genetic element

from one site to another in a DNA molecule [5]. This type of repeats recurs at many dispersed positions within a genome.

By types of their transpositional intermediates, interspersed repeats can be classified into two classes: RNA transposons and DNA transposons. As we do not consider interspersed repeats in this thesis, the interested audience can find a nice introduction of interspersed repeats in [5].

1.2 Research Significance

In recent years, researches have suggested that tandem repeats play a more and more important role in genetic markers, gene regulation, and human diseases.

Because the number of copies in any specific tandem repeat is often polymorphic in the population, some tandem repeats are used as the significant genetic markers for DNA fingerprinting, linkage analysis, and paternity testing [7, 12, 46]. For instance, in the United States, the core set of 13 tandem repeats are being used to generate the FBI Combined DNA Index System

(CODIS) database which has been successful at linking DNA profiles from crime scene evidence and at aiding paternity testing [31]. We also use 6 of these 13 tandem repeats to verify our algorithm in Chapter 6. Another example is that recent studies of allele diversity at tandem repeat loci have provided support for the "Out of Africa" hypothesis of modern human evolution [1, 42].

In addition, tandem repeats have been proven play a crucial role in gene regulation and appear to be linked with some types of transcription factor binding sites. They may interact with transcription factors where they are able to alter the structure of the chromatin or act as protein binding sites [26]. They are also found to have an apparent function in the development of immune system cells because it was found that breakpoints for immunoglobulin heavy chain switch recombination occur within tandem repeats preceding the heavy chain constant region genes [11].

Last but not least, the discovery of the trinucleotide repeat diseases and cancers has piqued interest in tandem repeats [35,

36, 40]. Those diseases, including Fridreich's ataxia (*GAA* repeats) [8], myotonic dystrophy (*CTG* repeats) [13], Huntington's disease (*CAG* repeats) [20], spinal and bullar muscular atrophy (*GAA* repeats) [37], and fragile-X mental retardation (*CGG* repeats) [44] are caused by the unexpectedly dramatic increase in the copy number of a trinucleotide pattern. In afflicted individuals, the copy number has been amplified from the normal range of tens of copies to hundreds or thousands [34]. It has been suggested that the repeats themselves produce unusual physical structures in the DNA, resulting in polymerase slippage and amplification [47, 48].

In conclusion, takeing tandem repeats for example, the detection of short adjacent repeats is of considerable significance.

1.3 Contributions

In this work, we make the following contributions to this repeats detection problem:

- Most existing methods have focused on the detection of

tandem repeats in a single DNA sequence, but none has the inherent ability to detect the common sequence pattern in multiple DNA sequences. We expand the scope of identifying tandem repeats to multiple DNA sequences, by relaxing the implicit assumption of a single DNA sequence in existing work. This is especially helpful in the case where tandem repeats shared by multiple DNA sequences from some particular loci shed light on molecular structure, function, and evolution. It can also be easily adapted to detect tandem repeats within a single DNA sequence. When the repeats in a single DNA sequence are spread around, the algorithm can be directly used by dividing the single sequence into multiple sub-sequences.

- We introduce a full probabilistic generative model for the approximate short adjacent repeats and a binary vector data structure to address the inter-unit insertions problem. The generative model in this thesis is well along the line of [21, 25], which used the sequence motif model to detect

the enriched pattern in multiple sequences. In the scenario of repeats detection, the same pattern is also enriched in a local area in each sequence. Our model is built to make use of the two levels of signal enrichment. We use Bayesian approach [14, 24] to infer the parameters of our model.

- We introduce a Bayesian approach to detect short adjacent repeats in a de novo fashion and use a collapsing technique [25] to improve computing efficiency. We demonstrate the effectiveness of the Markov chain Monte Carlo (MCMC) algorithm through experiments on both synthetic data and real biological data.

1.4 Thesis Organization

This thesis is divided into seven chapters and is organized as follow. Chapter 2 introduces existing methods for tandem repeats detection in the respect of repeats model. A brief overview of our method is also given in Chapter 2 and it helps audience understand our idea quickly. In Chapter 3, we go over preliminary

knowledge of multinomial distribution, Dirichlet distribution, Metropolis-Hastings sampling, and Gibbs sampling. In Chapter 4, we formulate the statistical problem, explore the parameter structure and deduce the posterior distribution of our model. The maximum a posteriori (MAP) estimate is used as the point estimate of the parameters. In Chapter 5, we formally present the schematic procedure of the MCMC algorithm and demonstrate it in detail step by step. In Chapter 6, we carry out extensive experiments on both synthetic data and real data to verify the algorithm. Chapter 7 concludes the thesis and issues some possible future works.

□ **End of chapter.**

Chapter 2

Literature Review and Overview of Our Method

Summary

In this chapter, we introduce existing methods for tandem repeats detection in the respect of repeats model: exact tandem repeats, fuzzy tandem repeats, and approximate tandem repeats. Then, a brief overview of our method is presented that helps audience understand our idea quickly.

2.1 Existing Methods

Most existing methods for identifying short adjacent repetitive patterns focus on tandem repeats in DNA sequence. Generally speaking, the methods of detecting tandem repeats in a single DNA sequence can be classified into three categories, depending on the tandem repeats models: exact tandem repeats (also called perfect tandem repeats), fuzzy tandem repeats and approximate tandem repeats.

The detection of exact tandem repeats is relatively easy and well studied. [27] described an algorithm (a variation of the Knuth-Morris-Pratt algorithm) that could find all exact tandem repeats in a sequence of length n, of which time complexity was $O(nlogn)$. [28] proposed an optimal seed-up parallel algorithm to detect tandem repeats. [9] was a program based on the Aho Corasick algorithm for finding exact tandem repeats using a keyword tree. [10] presented a vectorizable algorithm to find exact tandem repeats. [49] marked nucleotides of length 2 bp to 6 bp for different lists which were used to match the

sequence. However, since lots of substitutions (changing from one nucleotide to another, also called intra-unit mismatches in this thesis), insertions and deletions occur in DNA sequences, there are few exact tandem repeats existing in nature.

A number of methods for finding fuzzy tandem repeats, where the patterns differ from each other only by substitutions, were therefore developed. [4] focused on fuzzy tandem repeats with the pattern width ranging from 3 bp to 24 bp. Because the fuzzy tandem repeats require that the pattern width in the repeats is the same, this kind of algorithms are in vain when dealing with the other approximate tandem repeats. In addition to substitutions, there are insertions and deletions taking place in approximate tandem repeats.

In recent decades, many works focused on approximate tandem repeats detection. [30] presented a compression algorithm testing the presence of a particular type of dosDNA (approximate tandem repeats of small motifs with the pattern width no more than 4 bp). [32] proposed an algorithm that first filtered out non-repetitive regions of the sequence using statistical prop-

erties, and then detected all approximate tandem repeats which fulfilled certain criteria for the remaining parts. [2] also proposed the two-step method: it first searched for significant exact repetitions in the sequence and then used a threshold to check if these repetitions were actual approximate tandem repeats. Moreover, [2] developed a computer program called tandem repeats finder (TRF) based on this method.

All the above methods can be categorized as string matching algorithms. Recently, more and more signal processing methods have been proposed for addressing the repeats detection problem. [6] has designed an algorithm based on short time periodicity transform. [39] has used a tricolor spectrogram to show minisatellites in sequences. [43] has proposed a sum spectrum based on the Fourier transform. [33] has developed a computer program called the spectral repeats finder (SRF). [16] has presented an algorithm based on an orthogonal exact periodic subspace decomposition technique and also extended it to approximate tandem repeats. [50] has introduced a method based on parametric spectral estimation with two-step: it first analyzed

the spectrogram of a DNA sequence based on the autoregressive model, and then selected the significant peaks for searching tandem repeats within corresponding regions.

However, all those methods relied on the periodicity of a short segment in a single long sequence. Also, none has the inherent ability to detect the common tandem repeats in multiple DNA sequences.

2.2 Overview of Our Method

The mentioned tandem repeat example

$$\cdots ATCCGATCCGATCCG \cdots$$

shows an exact tandem repeat without any mismatches, insertions or deletions. In general, we represent an exact tandem repeat with pattern $x_1x_2\cdots x_J$ (the pattern width $J \geq 2$) and copy number Ω ($\Omega \geq 2$) as

$$X = (x_1x_2\cdots x_J)^{\Omega} = x_1x_2\cdots x_J \ldots x_1x_2\cdots x_J,$$

where $x_j, \forall j$ could be any of four letters A, T, C and G. For the sake of convenience, we name X as the repeat ***segment*** within

a DNA sequence, $x_1x_2\cdots x_J$ as the repeat ***unit*** which makes up a repeat segment and x_j as the j-th ***letter*** (also called ***nucleotide***) within a repeat unit. Compared with the exact tandem repeats, we are more concerned with the short adjacent repeats, which are approximate tandem repeats in the case of intra-unit mismatches and inter-unit insertions in this thesis. The inter-unit insertion denotes the case when one or more letters are inserted between any two adjacent repeat units.

An example of short adjacent repeats would be

$$\cdots ATCCGATgCGtATCCG\cdots$$

(the lower case highlights the variations), where the letter C is substituted by the letter g within the second repeat unit (the case of intra-unit mismatches), and a letter t is inserted between the second repeat unit and the third repeat unit (the case of inter-unit insertions).

In our work, we model the tandem repeats by a $4 \times J$ ***motif***

matrix [25] as

$$\Theta = \begin{bmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_J \end{bmatrix} = \begin{bmatrix} \theta_{A,1} & \theta_{A,2} & \cdots & \theta_{A,J} \\ \theta_{T,1} & \theta_{T,2} & \cdots & \theta_{T,J} \\ \theta_{C,1} & \theta_{C,2} & \cdots & \theta_{C,J} \\ \theta_{G,1} & \theta_{G,2} & \cdots & \theta_{G,J} \end{bmatrix},$$

where each column specifies the probabilities of finding the corresponding nucleotides in that position. Besides, we model the ***background*** area by using a vector as

$$\Phi = \begin{bmatrix} \phi_A & \phi_T & \phi_C & \phi_G \end{bmatrix}^T,$$

where each element represents the probability of finding the corresponding letter at a non-unit position. Noted that T denotes vector transpose.

Given a set of DNA sequences, each of which is embedded with a repeat segment composed of multiple repeat units (adjacent but allow inter-unit insertions in-between) generated from the same motif matrix, our goals are: (1) to detect where the repeat segment locates within each sequence; (2) to point out where the repeat units are within each repeat segment; (3) to

estimate the motif matrix that describes the pattern of the collection of all repeat units; (4) to estimate the distribution of the homogeneous background.

As far as we know, this is the first work using motif matrix [21, 25] for repeats detection. The key idea is to use the motif matrix to model repeat units as stochastic strings which are adjacently embedded in a homogeneous background, and to implement a Bayesian inference in the missing-data framework [23, 41]. We use $\mathbf{R}$, $\mathbf{A}$, and $\mathbf{S}$ to denote the set of given DNA sequences, the set of repeat segment starting positions and the set of within-segment structures, respectively. For the Bayesian inference of these parameters, we specify independent prior distributions for the parameters $\mathbf{\Theta}$, $\mathbf{A}$ and $\mathbf{S}$. A MCMC algorithm iterates the following sampling steps: $P(\Theta, \Phi|\mathbf{A}, \mathbf{S}, \mathbf{R})$, $P(\mathbf{A}|\mathbf{S}, \Theta, \Phi, \mathbf{R})$, and $P(\mathbf{S}|\mathbf{A}, \Theta, \Phi, \mathbf{R})$. Each of the sampling steps updates the corresponding parameter conditional on the current values of other parameters. After the MCMC iterations converge, the sampled parameter values give us a whole picture of their jointly posterior distribution.

□ **End of chapter.**

Chapter 3

Theoretical Background

Summary

In this chapter, we go over the preliminary knowledge used in Chapter 4 and 5. Multinomial distribution and Dirichlet distribution are introduced for better understanding of our generative model. Metropolis-Hastings sampling and Gibbs sampling are introduced for better understanding of the MCMC algorithm. Acknowledgement: some materials are from *Wikipedia*.

3.1 Multinomial Distributions

The multinomial distribution is a generalization of the binomial distribution. Suppose each trial result is exactly one of a fixed finite number N of possible outcomes with probabilities $p_1, ..., p_N$ and there are in total K independent trials, where $p_n \geq 0, 1 \leq n \leq N$ and $\sum_{n=1}^{N} p_n = 1$. Then let the discrete random variables X_n denote the number of times the n-th outcome was observed over the K trials. The vector $X = (X_1, ..., X_N)$ follows a multinomial distribution with parameters K and $(p_1, ..., p_N)$.

The probability mass function of the multinomial distribution is:

$$\begin{aligned} f(x_1, ..., x_N; K, p_1, ..., p_N) &= Pr(X_1 = x_1, ..., X_N = x_N) \\ &= \frac{\left(\sum_{n=1}^{N} x_n\right)!}{\prod_{n=1}^{N} (x_n!)} \prod_{n=1}^{N} p_n^{x_n}. \end{aligned} \tag{3.1}$$

3.2 Dirichlet Distribution

Similar to the relationship between the multinomial distribution and the binomial distribution, the Dirichlet distribution is a generalization of the beta distribution. Also it is conjugate prior

of the multinomial distribution in Bayesian statistics. That is, its probability density returns the belief that the probabilities of N rival events are x_n given each event has been observed $\alpha_n - 1$ times.

The Dirichlet distribution of order $N \geq 2$ with parameters $\alpha_n > 0, 1 \leq n \leq N$ has a probability density function:

$$f(x_1, ..., x_N; \alpha_1, ..., \alpha_N) = \frac{\Gamma\left(\sum_{n=1}^{N} \alpha_n\right)}{\prod_{n=1}^{N} \Gamma(\alpha_i)} \prod_{n=1}^{N} x_n^{\alpha_n - 1}, \qquad (3.2)$$

where Γ is the Gamma function, $x_n > 0, 1 \leq n \leq N$ and $\sum_{n=1}^{N} x_n = 1$.

An example of Dirichlet distribution can be given by considering an urn containing balls of N different colors. Initially, the urn contains α_1 balls of color 1, α_2 balls of color 2, and so on. Now perform K draws from the urn, where after each draw, the ball is placed back into the urn with an additional ball of the same color. In the limit as K approaches infinity, the proportions of different colored balls in the urn will be distributed as $Dirichlet(\alpha_1, ..., \alpha_N)$ [3].

3.3 Metropolis-Hastings Sampling

The Metropolis-Hastings algorithm [17, 29] is a MCMC method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult.

The Metropolis-Hastings algorithm can draw samples from any probability distribution $P(x)$, requiring only that a function proportional to the density can be calculated. In Bayesian applications, the normalization factor is often extremely difficult to compute, so the ability to generate a sample without knowing this constant of proportionality is a major virtue of the algorithm. The algorithm generates a Markov chain in which each state x^{t+1} depends only on the previous state x^t. The algorithm uses a proposal density $Q(x'; x^t)$, which depends on the current state x^t, to generate a new proposed sample x'. This proposal is accepted as the next value, i.e., $x^{t+1} = x'$ if α drawn from $Uniform(0, 1)$ satisfies

$$\alpha < min\left(1, \frac{P(x')\,Q(x^t; x')}{P(x^t)\,Q(x'; x^t)}\right). \tag{3.3}$$

Otherwise, the current value of x is retained: $x^{t+1} = x^t$.

3.4 Gibbs Sampling

The Gibbs sampling [15] is an algorithm to generate a sequence of samples from the joint probability of two or more random variables. It is a special case of the Metropolis-Hastings algorithm wherein the random value is always accepted and thus it is usually faster to use.

The key to the Gibbs sampling is that one only considers univariate conditional distributions. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms. For instance, consider a bivariate random variable (x, y), and suppose we wish to compute both marginals, $p(x)$ and $p(y)$. The idea behind the algorithm is that it is far easier to consider a sequence of conditional distributions, $p(x|y)$ and $p(y|x)$, than it is to obtain the marginal by integration of the joint density $p(x, y)$. The sampler starts with some initial value y_0 for y and obtains x_0 by generating a random variable from the conditional distribution $p(x|y = y_0)$. The sampler then uses x_0 to generate a new value of y_1, drawing from

the conditional distribution based on the value x_0, $p(y|x = x_0)$. The sampler proceeds as follows,

$$\begin{aligned} x_i &\sim p(x|y = y_{i-1}) \\ y_i &\sim p(y|x = x_i) \end{aligned} \tag{3.4}$$

Repeating this process k times, generates a Gibbs sequence of length k, where a subset of points (x_i, y_i) are taken as our simulated draws from the full joint distribution.

□ **End of chapter.**

Chapter 4

Problem Description

Summary

In this chapter, we present the input data and parameters of our generative model. Table 4.1 and Table 4.2 list the key notations used in this thesis. Then, as two evolving parameter groups are maintained, we investigate the relationship between the two parameter groups. Before concluding this chapter, we deduce the full posterior distribution as well as the collapsed posterior distribution.

4.1 Generative Model

In order to formulate our problem, we define four kinds of parameters corresponding to the four goals stated in the Section 2.2: (1) $\mathbf{A}$ is the set of repeat segment starting positions that reveals where the repeat segment locates within each sequence; (2) $\mathbf{S}$ is the set of repeat segment structures that could tell how many times the repeat unit repeats within each repeat segment and where they are; (3) Θ stands for the motif matrix that describes relative residue frequencies for each position of the repeat unit; (4) Φ represents the background distribution, i.e., the relative residue frequencies at a non-unit position.

In the following subsections, we give the exact definitions of the input data $\mathbf{R}$ and those four kinds of parameters $\mathbf{A}$, $\mathbf{S}$, Θ, and Φ, respectively. For ease of presentation, an example of the schematic diagram of our generative model is shown in the end of this section.

Table 4.1: Key notations I.

Notation	Definition
J	The pattern width.
$\mathbf{R}$	The set of all input DNA sequences.
R_n	The n-th sequence of $\mathbf{R}$.
$r_{n,l}$	The nucleotide at the l-th position in R_n.
$\sum$	The finite alphabet $\{A, T, C, G\}$ for DNA sequences.
L_n	The length of R_n.
$\mathbf{A}$	The set of repeat segment starting positions.
a_n	The repeat segment starting position of R_n.
$\mathbf{S}$	The set of repeat segment structures.
$\mathbf{s}_n$	The repeat segment structure of R_n.
$\xi_{n,z}$	The binary variable indicating if a repeat unit starts at the z-th position of $\mathbf{s}_n$ or not.
g_i	The gap length between the i-th repeat unit and the $i+1$-th repeat unit.

Table 4.2: Key notations II.

Notation	Definition
Z_n	The length of $\mathbf{s}_n$.
Ω	The copy number equal to $\lvert\mathbf{s}_n\rvert$.
G	The maximum allowed gap length.
Θ	The motif matrix.
$\boldsymbol{\theta}_j$	The probabilities of finding each letter at the j -th position of the repeat unit.
$\theta_{k,j}$	The probability of finding the letter k at the j -th position of the repeat unit.
Φ	The background distribution.
ϕ_k	The probability of finding the letter k at the non -unit position.

4.1.1 Input Data R

A set of N input sequences, denoted by $\mathbf{R}$, can be written as

$$\text{Data set } \mathbf{R}: \begin{matrix} \text{sequence } R_1: & r_{1,1} & r_{1,2} & \cdots & r_{1,L_1} \\ \text{sequence } R_2: & r_{2,1} & r_{2,2} & \cdots & r_{2,L_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{sequence } R_N: & r_{N,1} & r_{N,2} & \cdots & r_{N,L_N} \end{matrix},$$

where the residue $r_{n,l}, 1 \leq n \leq N, 1 \leq l \leq L_n$ is from the finite alphabet $\sum$ ($\sum = \{A, T, C, G\}$ for DNA sequences), and $L_n, 1 \leq n \leq N$ is the length of the n-th sequence R_n. Noted that it is not necessary that all sequences are of the same length. In the model, we assume that the multiple input sequences are mutually independent. We denote the collection of indices by $I = \{(n, l) : n = 1, \ldots, N; l = 1, \ldots, L_n\}$. For any set $W \subseteq I$, we define $\mathbf{R}_W = \{r_{n,l} : (n, l) \in W\}$. We also define $W^C = I \setminus W$ as the set of all elements which are members of I but not members of W.

4.1.2 Parameters A (Repeat Segment Starting Positions)

A set of repeat segment starting positions, denoted by $\mathbf{A}$, can be written as $\begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}^T$. It reveals where the repeat segment locates within each sequence. To begin with the simplest case, we assume that there is only one repeat segment per sequence. Similar to motif discovery problem [25], the algorithm can be extended to allow multiple repeat segments per sequence.

For each $P(a_n)$, we use a uniform distribution, which means each sequence is priorly considered to contain a repeat segment starting at a random position. We further assume the independence among all elements in $\mathbf{A}$. Thus,

$$P(\mathbf{A}) = \prod_{n=1}^{N} P(a_n) \propto 1. \tag{4.1}$$

4.1.3 Parameters S (Repeat Segment Structures)

A set of repeat segment structures, denoted by $\mathbf{S}$, can be written as $\begin{bmatrix} \mathbf{s}_1^T & \mathbf{s}_2^T & \cdots & \mathbf{s}_N^T \end{bmatrix}^T$, where $\mathbf{s}_n, 1 \leq n \leq N$ is a binary vector of the following format:

$$\mathbf{s}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & \xi_{n,J+g_1+1} & \cdots & \xi_{n,z} & \cdots & \xi_{n,Z_n} \end{bmatrix}.$$

The binary variable $\xi_{n,z} = 1$ indicates that there is a repeat unit occurring from z to $z + J - 1$ within the repeat segment $\mathbf{s}_n$, or equivalently, from $a_n + z - 1$ to $a_n + z + J - 2$ within the sequence R_n. To avoid non-identifiability, we require that the first element $\xi_{n,1}, 1 \leq n \leq N$ must be 1. The total number of 1 within $\mathbf{s}_n$, denoted by $|\mathbf{s}_n|$, is equal to the copy number Ω of the repeat segment. $g_i, 1 \leq i \leq \Omega - 1$ is the gap length between

the i-th repeat unit and the $i+1$-th repeat unit. We assume that each repeat segment is composed of multiple repeat units separated by gaps of random length from 0 to G, where G is the maximum allowed gap length. By introducing the binary vector $\mathbf{s}_n$, we are now able to deal with the case of inter-unit insertions.

For $P(\mathbf{s}_n)$, we assume it is only determined by the copy number, in other words, the number of 1 in the vector $\mathbf{s}_n$. More specifically, we set $P(\mathbf{s}_n) \propto \varepsilon^{|\mathbf{s}_n|}$. The tuning constant ε is a positive real number less than 1, which means the probability will decrease if there are more repeat units. Without loss of generality, we make the assumption that all elements in $\mathbf{S}$ are mutually independent. Thus,

$$P(\mathbf{S}) = \prod_{n=1}^{N} P(\mathbf{s}_n) \propto \varepsilon^{\sum_{n=1}^{N} |\mathbf{s}_n|}. \tag{4.2}$$

4.1.4 Parameters Θ (Motif Matrix)

The motif matrix denoted by Θ, can be written as a $4 \times J$ motif matrix [25] as

$$\Theta = \begin{bmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_J \end{bmatrix} = \begin{bmatrix} \theta_{A,1} & \theta_{A,2} & \cdots & \theta_{A,J} \\ \theta_{T,1} & \theta_{T,2} & \cdots & \theta_{T,J} \\ \theta_{C,1} & \theta_{C,2} & \cdots & \theta_{C,J} \\ \theta_{G,1} & \theta_{G,2} & \cdots & \theta_{G,J} \end{bmatrix},$$

where $\theta_{k,j}, k \in \sum, 1 \leq j \leq J$ is the probability of finding the letter k at the position j among all repeat units.

It is assumed that all repeat units are independent and identical samples from this motif matrix Θ of width J. We take Θ as the product multinomial model with product Dirichlet priors $\mathbf{B}$ [25] as,

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_J \end{bmatrix} = \begin{bmatrix} \beta_{A,1} & \beta_{A,2} & \cdots & \beta_{A,J} \\ \beta_{T,1} & \beta_{T,2} & \cdots & \beta_{T,J} \\ \beta_{C,1} & \beta_{C,2} & \cdots & \beta_{C,J} \\ \beta_{G,1} & \beta_{G,2} & \cdots & \beta_{G,J} \end{bmatrix},$$

if each $\boldsymbol{\theta}_j$ are independent and 4-dimensional Dirichlet random variables with distributions $Dirichlet\left(\boldsymbol{\beta}_j\right)$. Without any prior

knowledge, we set all elements in $\mathbf{B}$ equal to 1. As a result, we have

$$P(\Theta) \propto 1. \tag{4.3}$$

4.1.5 Parameters Φ (Background Distribution)

The background distribution denoted by Φ, can be written as

$$\Phi = \begin{bmatrix} \phi_A & \phi_T & \phi_C & \phi_G \end{bmatrix}^T,$$

, where $\phi_k, k \in \sum$ is the probabilities of finding each letter k at a non-unit position.

By analogy with the subsection 4.1.4, we assume that all nucleotides at non-unit positions are independent and identical samples from the background distribution Φ. Similar to Θ, Φ represents a multinomial model with Dirichlet priors $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_A & \alpha_T & \alpha_C & \alpha_G \end{bmatrix}^T$. Also, we set all elements in $\boldsymbol{\alpha}$ equal to 1. Consequently, we have

$$P(\Phi) \propto 1. \tag{4.4}$$

4.1.6 An Example of the Model Schematic Diagram

For ease of presentation, Figure 4.1 shows an example of the schematic diagram of our repeats model. In this specific model, there are 5 sequences with each length equal to $L_n, 1 \leq n \leq 5$. For each repeat segment, the starting position and structure are shown on top of the gray area. The background area is painted in white, whatever the borderline is dotted or solid. Each white square with solid borderline represents a gap with length 1.

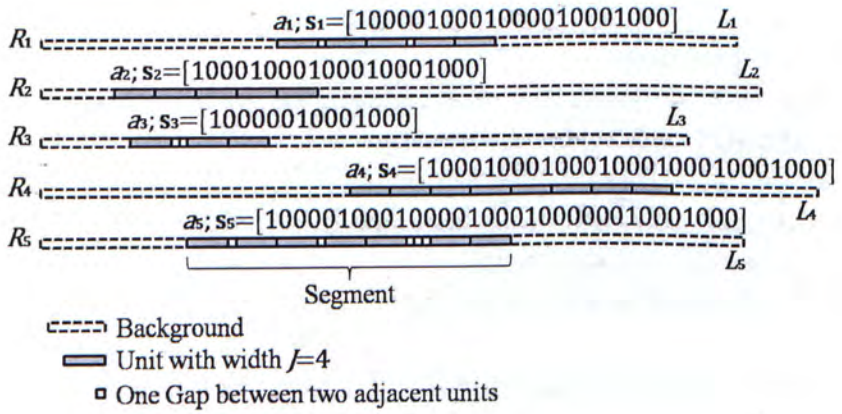

Figure 4.1: The schematic diagram of the generative model under the settings $N = 5$ and $J = 4$.

4.2 Parameter Structure

The algorithm maintains two evolving parameter groups. The first group, **A** and **S**, indicates the locations of all repeat units. The second group, Θ and Φ, describes the repetitive pattern and the background distribution.

Before investigating the relationship between the two parameter groups, we introduce a counting function **h** [25]. For a given set of categorical data, e.g., $Y = \{y_1, \cdots, y_l, \cdots, y_L\}$, where each y_l takes values from $\sum$, we define the counting function **h** such that $\mathbf{h}(Y) = [\ m_A \ \ m_T \ \ m_C \ \ m_G\ ]^T$, where $m_k, k \in \sum$ is the total number of letter k observed in Y. It is noted that the function **h** has an additive property. If another data set Y' is given, then $\mathbf{h}(Y) + \mathbf{h}(Y') = \mathbf{h}(Y \oplus Y')$, where the left side is just the ordinary addition for vectors and $Y \oplus Y'$ indicates combining the two categorical data sets Y and Y'.

Given **A** and **S**, the observation of all repeat units' indices can be denoted as set

$$U = \{(n, \xi_{n,z}(a_n + z - 1 + j - 1))\},$$

where $n = 1, \ldots, N, z = 1, \ldots, Z_n, j = 1, ...J$. Notice that the element $(n, \xi_{n,z}(a_n + z - 1 + j - 1))$ exists if and only if $\xi_{n,z} = 1$, because there is no 0-th position in the sequences. As U gives the indices of all repeat units, the observations of all letters at non-unit positions can be denoted as $\mathbf{R}_{U^C}$. We also write the set of the residues occupied in the j-th positions of all repeat units as

$$U(j) = \{(n, \xi_{n,z}(a_n + z - 1 + j - 1))\},$$

where $n = 1, \ldots, N, z = 1, \ldots, Z_n$.

We can get the sufficient statistics by applying the counting function $\mathbf{h}$ on $\mathbf{R}_{U^C}$ and $\mathbf{R}_{U(j)}, 1 \leq j \leq J$. We denote the results as a vector $\mathbf{h}(\mathbf{R}_{U^C}) = \begin{bmatrix} w_A & w_T & w_C & w_G \end{bmatrix}^T$ and a $4 \times J$ matrix

$$\mathbf{h}(\mathbf{R}_U) = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_J \end{bmatrix} = \begin{bmatrix} m_{A,1} & m_{A,2} & \cdots & m_{A,J} \\ m_{T,1} & m_{T,2} & \cdots & m_{T,J} \\ m_{C,1} & m_{C,2} & \cdots & m_{C,J} \\ m_{G,1} & m_{G,2} & \cdots & m_{G,J} \end{bmatrix}.$$

As the same assumption as [25], $\mathbf{h}(\mathbf{R}_U)$ follows a product multi-

nomial distribution (PM) with parameter Θ, i.e.,

$$\mathbf{h}(\mathbf{R}_U) \sim PM(\Theta; |\mathbf{m}_1|, |\mathbf{m}_2|, \cdots, |\mathbf{m}_J|),$$

where $|\mathbf{m}_j| = m_{A,j} + m_{T,j} + m_{C,j} + m_{G,j}, 1 \leq j \leq J$, if each $\mathbf{m}_j$ follows the multinomial distribution $Multinomial(\boldsymbol{\theta}_j, |\mathbf{m}_j|)$. $\mathbf{h}(\mathbf{R}_{U^C})$ follows a multinomial distribution with parameter Φ, i.e.,

$$\mathbf{h}(\mathbf{R}_{U^C}) \sim Multinomial(\Phi; w_A + w_T + w_C + w_G).$$

For the Bayesian inference of Θ and Φ, we use conjugate priors, which are product Dirichlet distribution (PD) [25] and Dirichlet distribution, respectively. Therefore, the posterior distribution of Θ is $PD(\mathbf{B} + \mathbf{h}(\mathbf{R}_U))$, if each $\boldsymbol{\theta}_j$ are independent and 4-dimensional Dirichlet random variables with distributions $Dirichlet(\boldsymbol{\beta}_j)$. Similarly, the posterior distribution of Φ is $Dirichlet(\boldsymbol{\alpha} + \mathbf{h}(\mathbf{R}_{U^C}))$.

4.3 Posterior Distribution

In this section, we first present the full posterior distribution. Then, a collapsing technique is used to make the proposed al-

gorithm introduced in the next chapter computationally more efficient.

4.3.1 The Full Posterior Distribution

Given $\mathbf{A}$, $\mathbf{S}$, Θ and Φ, the probability of observing the given data $\mathbf{R}$ can be written as

$$P(\mathbf{R}|\mathbf{A},\mathbf{S},\Theta,\Phi) = \Phi^{\mathbf{h}(\mathbf{R}_{UC})} \prod_{j=1}^{J} \boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{U(j)})}, \tag{4.5}$$

where we define the vector power of a vector as the product of all elements after taking corresponding power, i.e., $\Phi^{\mathbf{h}(\mathbf{R}_{UC})} = \phi_A^{w_A}\phi_T^{w_T}\phi_C^{w_C}\phi_G^{w_G}$ and $\boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{U(j)})} = \theta_{A,j}^{m_{A,j}}\theta_{T,j}^{m_{T,j}}\theta_{C,j}^{m_{C,j}}\theta_{G,j}^{m_{G,j}}$. With mutually independent priors, the jointly posterior distribution of $\mathbf{A}$, $\mathbf{S}$, Θ and Φ can be written as

$$P(\mathbf{A},\mathbf{S},\Theta,\Phi|\mathbf{R}) \propto P(\mathbf{R}|\mathbf{A},\mathbf{S},\Theta,\Phi)P(\mathbf{A})P(\mathbf{S})P(\Theta)P(\Phi).$$

Due to the complete-data likelihood of all parameters: Equation 4.5 and the prior distributions of each kind of parameters: Equation 4.1, Equation 4.2, Equation 4.3, and Equation 4.4, we

can write the full posterior distribution as

$$P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R}) \propto \varepsilon^{\sum_{n=1}^{N}|\mathbf{s}_n|}\Phi^{\mathbf{h}(\mathbf{R}_{UC})}\prod_{j=1}^{J}\boldsymbol{\theta}_j^{\mathbf{h}(\mathbf{R}_{U(j)})}. \quad (4.6)$$

4.3.2 The Collapsed Posterior Distribution

The MCMC algorithm based on the above jointly posterior distribution can be low efficient because of the big dimension of the solution / search space as well as the product Dirichlet sampler for Θ. A widely used solution is to integrate out the nuisance parameters. As $\mathbf{A}$ and $\mathbf{S}$ are mainly concerned, Θ and Φ are thus nuisance parameters. Or at least, it is not difficult to estimate approximate Θ or Φ given $\mathbf{A}$ and $\mathbf{S}$. Using the collapsing technique of [25], we can actually integrate out Θ and Φ in order to make the proposed algorithm computationally more efficient.

Noted that

$$P(\mathbf{A}, \mathbf{S}|\mathbf{R}) = \int\int P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R})d\Theta d\Phi,$$

our choices of the Dirichlet priors for Θ and Φ enable us to integrate out both Θ and Φ analytically. The resulting collapsed

posterior distribution can be written as

$$\begin{aligned} & P(\mathbf{A}, \mathbf{S}|\mathbf{R}) \\ & \propto \varepsilon^{\sum_{n=1}^{N} |\mathbf{s}_n|} \Gamma\left(\mathbf{h}\left(\mathbf{R}_{U^C}\right) + \boldsymbol{\alpha}\right) \prod_{j=1}^{J} \Gamma\left(\mathbf{h}\left(\mathbf{R}_{U(j)}\right) + \boldsymbol{\beta}_j\right). \end{aligned} \tag{4.7}$$

4.4 Conclusion

In this chapter, we introduce our generative model with the input data and four kinds of parameters which can be classified into two groups. Our probabilistic model makes the following assumptions: (1) the multiple input sequences are mutually independent; (2) each sequence contains a repeat segment starting at a random position; (3) each repeat segment is composed of multiple repeat units separated by gaps of random length from 0 bp to G bp; (4) all repeat units are independent and identical samples from the same motif matrix Θ of width J; (5) all nucleotides at non-unit positions are independent and identical samples from the background distribution Φ.

Moreover, we make the relationship between the two parameters groups explicit: the posterior distribution of Θ follows

$PD\left(\mathbf{B}+\mathbf{h}\left(\mathbf{R}_U\right)\right)$ and the posterior distribution of Φ follows $Dirichlet\left(\boldsymbol{\alpha}+\mathbf{h}\left(\mathbf{R}_{U^C}\right)\right)$, where $\mathbf{R}_U$ and $\mathbf{R}_{U^C}$ are determined by $\mathbf{A}$ and $\mathbf{S}$.

Last, we deduce both the full and collapsed posterior distribution. Our goal is to characterize the parameter values (both point estimate and uncertainty) by simulating from the posterior distribution.

□ **End of chapter.**

Chapter 5

Methodology

Summary

At the beginning of this chapter, we present the schematic procedure as shown in Table 5.1 and Table 5.2, respectively. Then, we present an elaborate algorithm based on the improved one which is shown in Table 5.2 step by step.

5.1 Schematic Procedure

5.1.1 The Basic Schematic Procedure

Given a set of N sequences $\mathbf{R}$, our task is to seek one repeat segment per sequence. All the N repeat segments consist of the same or alike repeat unit with equivalent width J. The algorithm maintains two groups of evolving parameter structures: (1) $\mathbf{A}$ and $\mathbf{S}$, which indicate the set of all repeat units $\mathbf{R}_U$, and (2) Θ and Φ, which describe the motif matrix and the background distribution. Our objective is to explore the most probable repeat segment location and structure within each sequence. These N repeat segments are obtained by locating $\mathbf{A}$ and adjusting $\mathbf{S}$ to maximize the posterior probability of the matching area $\mathbf{R}_U$. In other words, the target is to find out the maximum a posteriori (MAP) of $P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R})$.

The basic idea of addressing the optimization problem is to use Metropolis-in-Gibbs scheme [14, 24] to fully explore the posterior distribution, as shown in Table 5.1.

Table 5.1: The basic schematic procedure.

Step 1:	Initialize $\mathbf{A}$, $\mathbf{S}$, Θ and Φ;
Step 2:	Sample and Update $\mathbf{A}$ via $P(\mathbf{A}\|\mathbf{S}, \Theta, \Phi, \mathbf{R})$;
Step 3:	Sample and Update $\mathbf{S}$ via $P(\mathbf{S}\|\mathbf{A}, \Theta, \Phi, \mathbf{R})$;
Step 4:	Sample and Update Θ and Φ via $P(\Theta, \Phi\|\mathbf{A}, \mathbf{S}, \mathbf{R})$;
Step 5:	Repeat Step 2-4 until convergence;

5.1.2 The Improved Schematic Procedure

However, this standard Metropolis-in-Gibbs scheme based on the full posterior distribution is too time-consuming because it involves sampling from a product Dirichlet distribution [25]. Instead, we work on the collapsed posterior distribution $P(\mathbf{A}, \mathbf{S}|\mathbf{R})$ to improve computing efficiency. The idea is to integrate out the nuisance parameters Θ and Φ. The improved schematic procedure is shown in Table 5.2.

After initialization, the MCMC algorithm proceeds through iterations, each of which updates a_n and $\mathbf{s}_n$ one sequence after another in ascending order from 1 to N or at random. Within each iterative procedure, we pretend that $N-1$ repeat segments have been known, and we stochastically predict for the repeat

segment within the remaining sequence. More specifically, when the n-th sequence R_n is selected, we use the given information, $\mathbf{A}_{[-n]}$ and $\mathbf{S}_{[-n]}$, to estimate the current 'motif matrix' $\hat{\Theta}_n$ and 'background distribution' $\hat{\Phi}_n$ so as to determine new a_n and $\mathbf{s}_n$ sequentially. Here, $\mathbf{A}_{[-n]}$ denotes the set

$$\begin{bmatrix} a_1 & \cdots & a_{n-1} & a_{n+1} & \cdots & a_N \end{bmatrix}^T,$$

and $\mathbf{S}_{[-n]}$ denotes the set

$$\begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_{n-1} & \mathbf{s}_{n+1} & \cdots & \mathbf{s}_N \end{bmatrix}^T.$$

Intuitively, the more accurate the estimated motif matrix $\hat{\Theta}_n$ and background distribution $\hat{\Phi}_n$ constructed in the predictive update step, the more accurate the determination of a_n and $\mathbf{s}_n$ in the following sampling steps, and vice versa.

Table 5.2: The improved schematic procedure.

Step 1:	Initialize $\mathbf{A}$ and $\mathbf{S}$;
Step 2:	for n from 1 to N do 2.1: Calculate $\hat{\Theta}_n$ and $\hat{\Phi}_n$ via $\mathbf{A}_{[-n]}$ and $\mathbf{S}_{[-n]}$; 2.2: Sample and Update a_n via $P(a_n \mid \mathbf{s}_n, \hat{\Theta}_n, \hat{\Phi}_n, \mathbf{R})$; 2.3: Sample and Update $\mathbf{s}_n$ via $P(\mathbf{s}_n \mid a_n, \hat{\Theta}_n, \hat{\Phi}_n, \mathbf{R})$;
Step 3:	Repeat Step 2 until convergence;

5.2 Initialization

As shown in Table 5.2, the first step is to initialize $\mathbf{A}$ and $\mathbf{S}$. $\mathbf{A}$ is a $N \times 1$ column vector and the initial value of $a_n, 1 \leq n \leq N$ is randomly chosen from all of its possible values $1, \ldots, L_n - J + 1$. $\mathbf{S}$ is a $N \times Z$ binary matrix and all its row vectors are initially set as $\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$, i.e., all elements are 0 except the first one.

5.3 Predictive Update Step for $\hat{\Theta}_n$ and $\hat{\Phi}_n$

Suppose we are proceeding through the i-th iteration, the estimated $\hat{\Theta}_n$ is calculated as,

$$\hat{\boldsymbol{\theta}}_{n\cdot j}^{(i)} = \frac{\boldsymbol{\beta}_j + \mathbf{h}\left(\mathbf{R}_{U_{[-n]}^{(i-1)}(j)}\right)}{|\boldsymbol{\beta}_j| + \left|\mathbf{h}\left(\mathbf{R}_{U_{[-n]}^{(i-1)}(j)}\right)\right|}, j = 1, \ldots, J, \tag{5.1}$$

where $U_{[-n]}$ denotes all repeat units' indices excluding those repeat units' within R_n. $\hat{\boldsymbol{\theta}}_{n\cdot j}^{(i)}$ is the j-th column vector in $\hat{\Theta}_n^{(i)}$. The estimated $\hat{\Phi}_n$ is calculated as,

$$\hat{\Phi}_n^{(i)} = \frac{\boldsymbol{\alpha} + \mathbf{h}\left(\mathbf{R}_{[-n]}\right) - \sum_{j=1}^{J} \mathbf{h}\left(\mathbf{R}_{U_{[-n]}^{(i-1)}(j)}\right)}{|\boldsymbol{\alpha}| + \left|\mathbf{h}\left(\mathbf{R}_{[-n]}\right)\right| - \sum_{j=1}^{J} \left|\mathbf{h}\left(\mathbf{R}_{U_{[-n]}^{(i-1)}(j)}\right)\right|}. \tag{5.2}$$

Noted that we define the absolute value of a vector as the summation of all elements within the vector.

5.4 Gibbs Sampling Step for a_n

On condition that the repeat segment structure within the n-th sequence is known, we consider every possible repeat segment X with this structure as a promising instance. Using the results obtained in the predictive update step, we calculate the

probability of generating those matching repeat units within X according to the current 'motif matrix' $\hat{\Theta}_n^{(i)}$ and the probability of generating all letters within X according to the 'background distribution' $\hat{\Phi}_n^{(i)}$, respectively. The ratio of these two probabilities is assigned as the weight to each X, of which starting position is from 1 to $L_n - J + 1$. To sum up, we use the Gibbs sampling to update new $a_n^{(i)} = a'_n$ as follows,

$$P(a'_n|\mathbf{s}_n^{(i-1)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R}) \propto \prod_{j=1}^{J} \prod_{z=1}^{Z} \xi_{n,z}^{(i-1)} \frac{\hat{\theta}^{(i)}_{R_n(a_n+z-1+j-1),j}}{\hat{\phi}^{(i)}_{R_n(a_n+z-1+j-1)}}. \quad (5.3)$$

Notice that the component equals to 1 if $\xi_{n,z} = 0$. The normalized factor is the summation of all instances, i.e.,

$$\sum_{a'_n=1}^{L_n-J+1} P(a'_n|\mathbf{s}_n^{(i-1)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R}).$$

5.5 Metropolis-Hastings Sampling Step for s_n

The $1 \times Z$ binary vector $\mathbf{s}_n$ under the settings J and G allows at least $(G+1)^{\lfloor \frac{Z-J+1}{J+G} \rfloor}$ valid states, where $\lfloor\rfloor$ is the floor function. As the dimension of $\mathbf{s}_n$, which is Z, is usually a very large number, it is extremely difficult to compute the normalization constant of $P(\mathbf{s}_n|a_n, \hat{\Theta}_n, \hat{\Phi}_n, \mathbf{R})$. Based on that fact, we use the

Metropolis-Hastings sampling to update new $\mathbf{s}_n^{(i)}$ because it has the ability to generate a sample without knowing this constant.

In our case, the Hastings ratio can be written as

$$\lambda = \frac{P(\mathbf{s}'_n|a_n^{(i)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R})P\left(\mathbf{s}_n^{(i-1)}; \mathbf{s}'_n\right)}{P(\mathbf{s}_n^{(i-1)}|a_n^{(i)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R})P\left(\mathbf{s}'_n; \mathbf{s}_n^{(i-1)}\right)}, \tag{5.4}$$

where $P\left(\mathbf{s}'_n; \mathbf{s}_n^{(i-1)}\right)$ is the proposal density, which specifies the probability of proposing a move to $\mathbf{s}'_n$ given the previous state $\mathbf{s}_n^{(i-1)}$. The move is accepted, $\mathbf{s}_n^{(i)} = \mathbf{s}'_n$, with the probability $min\,(1, \lambda)$; otherwise, $\mathbf{s}_n^{(i)} = \mathbf{s}_n^{(i-1)}$.

In order to make the Markov chain ergodic and converge fast, we design five moves which can be categorized into three types as shown in Figure 5.1. For the sake of convenience, we transform the binary vector $\mathbf{s}_n$ into another format,

$$\left[\begin{array}{ccccccc} g_1 & \cdots & g_i & \cdots & g_{\Omega-1} \end{array}\right],$$

where g_i is the gap length between the i-th repeat unit and the $i+1$-th repeat unit and Ω is the copy number, i.e., $\Omega = |\mathbf{s}_n|$. We denote $Q_i, 1 \leq i \leq 5$ as each transition probability from the current state to the corresponding state next time. They satisfy

the constraint $\sum_{i=1}^{5} Q_i = 1$. An example of full state transition diagram under the settings $G = 2$ and $\Omega_{max} = 4$ is shown in Figure 5.2.

5.5.1 Rear Indel Move

The first type of move is to insert a repeat unit from all its possible positions behind the current last repeat unit with the transition probability Q_1 or delete the current last repeat unit with the transition probability Q_2. Let z_n^b be the starting position of the b-th repeat unit within the repeat segment $\mathbf{s}_n$, and it is not hard to say, $z_n^{|\mathbf{s}_n|}$ is the starting position of last repeat unit within $\mathbf{s}_n$. For the rear insertion case, the possible position ν is chosen from the range $a_n^{(i)} + z_n^{|\mathbf{s}_n^{(i-1)}|} + J - 1$ to $a_n^{(i)} + z_n^{|\mathbf{s}_n^{(i-1)}|} + J + G - 1$. Therefore, we write the Hastings ratio as follows,

$$\lambda = \varepsilon \frac{\prod_{j=1}^{J} \hat{\theta}^{(i)}_{r_{n,\nu+j-1},j}}{\prod_{j=1}^{J} \hat{\phi}^{(i)}_{r_{n,\nu+j-1}}} \frac{Q_2}{Q_1/(G+1)}. \tag{5.5}$$

For the rear deletion case, as we know the last repeat unit within the repeat segment is located at $\nu = a_n^{(i)} + z_n^{|\mathbf{s}_n^{(i-1)}|} - 1$. Thus, we

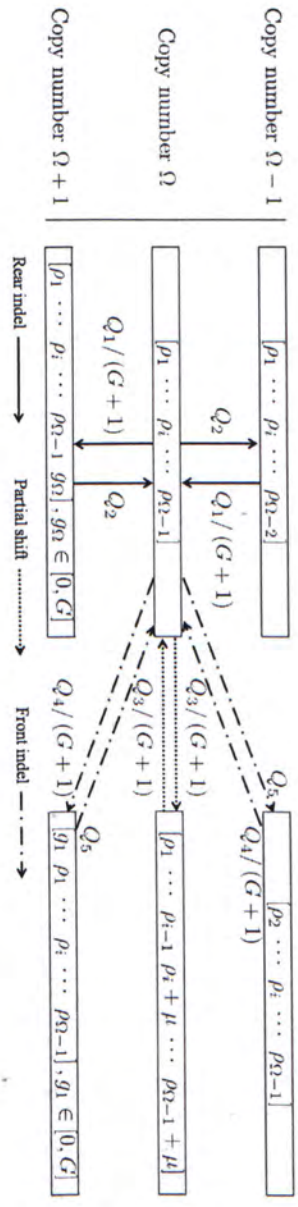

Figure 5.1: State transition diagram of the current state $\mathbf{s}_n$.

ρ_i is the deterministic gap length and g_i is the candidate gap length randomly chosen from the range 0 to G between the i-th repeat unit and the $i+1$-th repeat unit. $Q_i, 1 \leq i \leq 5$ is the transition probability for each move under three categories: rear indel move, partial shift move, and front indel move.

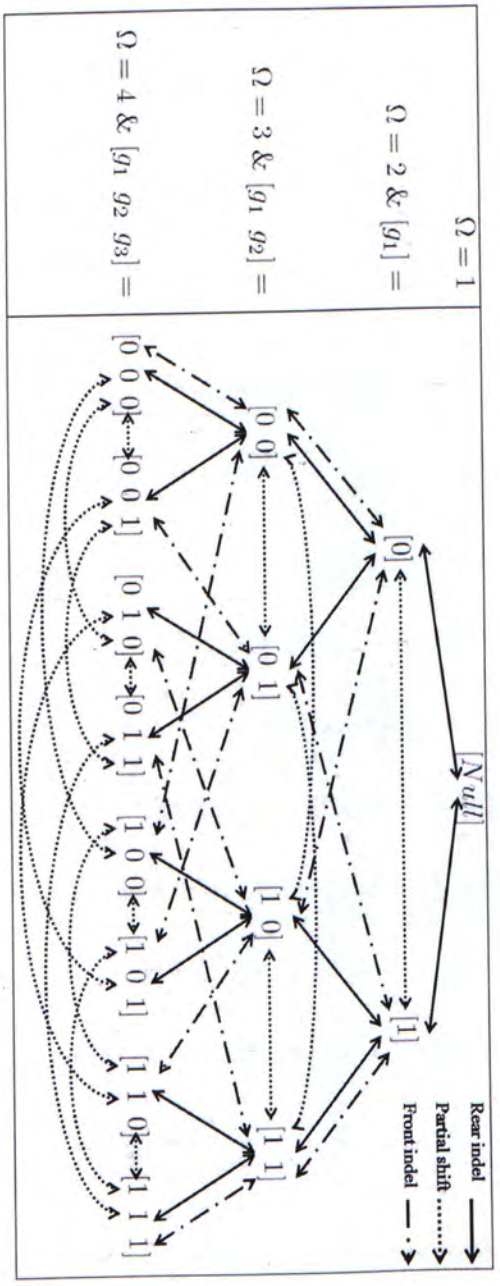

Figure 5.2: The full state transition diagram under the settings $G = 2$ and $\Omega_{max} = 4$.

have

$$\lambda = \frac{1}{\varepsilon} \frac{\prod_{j=1}^{J} \hat{\phi}^{(i)}_{r_{n,\nu+j-1}}}{\prod_{j=1}^{J} \hat{\theta}^{(i)}_{r_{n,\nu+j-1},j}} \frac{Q_1/(G+1)}{Q_2}. \tag{5.6}$$

5.5.2 Partial Shift Move

This move is to make partial shifts within the repeat segment $\mathbf{s}_n$ and its transition probability is Q_3. We randomly choose the order b in the range of $\left[2, |\mathbf{s}_n^{(i-1)}|\right]$, and we make partial shifts from the position z_n^b. The number of shifts μ is also randomly chosen among all its possible values from $z_n^{b-1} - z_n^b + J$ to $z_n^{b-1} - z_n^b + J + G$. Thus we have

$$\lambda = \frac{\prod_{j=1}^{J} \prod_{z=z_n^b}^{|\mathbf{s}_n^{(i-1)}|+J-1} \xi_{n,z}^{(i-1)} \frac{\hat{\theta}^{(i)}_{r_{n,a_n^{(i)}+z-1+j-1+\mu},j}}{\hat{\phi}^{(i)}_{r_{n,a_n^{(i)}+z-1+j-1+\mu}}}}{\prod_{j=1}^{J} \prod_{z=z_n^b}^{|\mathbf{s}_n^{(i-1)}|+J-1} \xi_{n,z}^{(i-1)} \frac{\hat{\theta}^{(i)}_{r_{n,a_n^{(i)}+z-1+j-1},j}}{\hat{\phi}^{(i)}_{r_{n,a_n^{(i)}+z-1+j-1}}}} \frac{Q_3/(G+1)}{Q_3/(G+1)}. \tag{5.7}$$

Notice that the component equals to 1 if $\xi_{n,z} = 0$.

5.5.3 Front Indel Move

Actually, the last type is to move $\mathbf{s}_n$ and corresponding a_n as a group using Metropolis-Hastings algorithm. For the sake of clear organization, we do not plan to separate it into another

subsection. In according with the rear indel move, we design the inserting move, with transition probability Q_4, as adding a repeat unit from a possible position $\nu \in \left[a_n^{(i)} - G - J, a_n^{(i)} - J\right]$ in front of the current repeat segment starting position. If the move is accepted, we renew $a_n^{(i)}$ with ν at the same time. Also, we design the deletion move, with the transition probability Q_5, as removing the first repeat unit within the repeat segment while renewing $a_n^{(i)}$ with $a_n^{(i)} + z_n^2 - 1$ if accepted. The corresponding Hastings ratios are

$$\lambda = \varepsilon \frac{\prod_{j=1}^{J} \hat{\theta}^{(i)}_{r_{n,\nu+j-1},j}}{\prod_{j=1}^{J} \hat{\phi}^{(i)}_{r_{n,\nu+j-1}}} \frac{Q_5}{Q_4/(G+1)} \tag{5.8}$$

and

$$\lambda = \frac{1}{\varepsilon} \frac{\prod_{j=1}^{J} \hat{\phi}^{(i)}_{r_{n,a_n^{(i)}+j-1}}}{\prod_{j=1}^{J} \hat{\theta}^{(i)}_{r_{n,a_n^{(i)}+j-1},j}} \frac{Q_4/(G+1)}{Q_5}, \tag{5.9}$$

respectively.

5.6 Phase Shifts

Although the collapsed sampler seems to work well in the MCMC algorithm, it may face the *phase* problems [21], which gets the MCMC algorithm stuck in a local optimum.

An example of the phase problem can be illustrated as follows. Suppose it is given 4 sequences, each of which repeat segment consists of perfect repeat units without gaps, and starts at the position 100, 200, 300 and 400, respectively. The pattern width is 20 bp. If the current iteration has $a_1 = 97$, $a_2 = 197$ and $a_3 = 297$, it will most likely proceed to choose $a_4 = 397$. Local moves are incapable of escaping this local optimum.

The solution is to compare the current $\mathbf{A}$ with sets shifted left and right [21]. We randomly choose a reasonable number μ and denote a $1 \times N$ vector $\boldsymbol{\mu} = \begin{bmatrix} \mu & \cdots & \mu \end{bmatrix}^T$. Then, the Hastings ratio will be

$$\lambda = \frac{P(\mathbf{A} + \boldsymbol{\mu}|\mathbf{R}, \mathbf{S})}{P(\mathbf{A}|\mathbf{R}, \mathbf{S})} \propto \frac{P(\mathbf{R}|\mathbf{A} + \boldsymbol{\mu}, \mathbf{S})}{P(\mathbf{R}|\mathbf{A}, \mathbf{S})}. \tag{5.10}$$

5.7 Conclusion

In this chapter, we present both the basic schematic procedure based on the full posterior distribution and the improved schematic procedure based on the collapsed posterior distribution. Because the former one involves sampling from a product

Dirichlet distributions, resulting in time-consuming, we prefer to choose the latter one.

After initialization, the MCMC algorithm proceeds through iterations, each of which has three steps: predictive update step for $\hat{\Theta}_n$ and $\hat{\Phi}_n$, Gibbs sampling step for a_n, and Metropolis-Hastings sampling step for $\mathbf{s_n}$. Intuitively, the more accurate the estimated motif matrix $\hat{\Theta}_n$ and background distribution $\hat{\Phi}_n$ constructed in the predictive update step, the more accurate the determination of a_n and $\mathbf{s}_n$ in the following sampling steps, and vice versa.

Moreover, in case of the MCMC algorithm get stuck in a local optimum. We execute the phase shifts step after every M-th iterations.

□ **End of chapter.**

Chapter 6

Results and Discussion

Summary

In this chapter, we present the results of two experiments: one on synthetic data and the other on real data. The test on synthetic data helps us explore the effectiveness of our algorithm. Also, we show how to choose the pattern width J via this experiment. The test on real data, compared with the widely used program Tandem Repeats Finder [2], validates the algorithm in practice.

6.1 Settings

We design two experiments, one on synthetic data (Experiment 1) to explores how effective the algorithm is and the other on real data (Experiment 2) to shows its performance on real data. All tests are conducted using C++ on a PC with 2.66G CPU and 2G memory.

In Experiment 1, we use synthetic data for ease of carrying out the experiment and evaluating the performance in a fully controlled manner. We generate the synthetic DNA sequences set $\mathbf{R}$. It contains N sequences with equivalent length L, each of which contains only one repeat segment whose starting position is randomly selected. Within each repeat segment, we also randomly pick out the copy number Ω in the range of $[\Omega_{min}, \Omega_{max}]$ and the gap length $g_i, 1 \leq i \leq \Omega - 1$ between the i-th repeat unit and the $i+1$-th repeat unit in the range of $[0, G]$. All repeat units with width J are generated following the predetermined motif matrix Θ, and the background distribution is assumed to be $\Phi = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}^T$. With all parameters and ε

known, the actual value of the jointly posterior probability can be calculated for reference. We evaluate the performance based on: (1) whether the joint posterior probability converges to its actual value; (2) what the false positives (FP) and the false negatives (FN) (also called type I and II errors, respectively) are. The false positive is the number of predicted repeat units which are not actual repeat units. The false negative is the number of actual repeat units which are not predicted by our algorithm. We report an error whenever the address of the examining repeat unit in one set can not be found in the other set.

Since the algorithm requires the pattern width J to be input, we also demonstrate how to select J via Experiment 1. The method is to execute the algorithm with a range of plausible widths and then choose the best result according to the MAP. Noted that the solution space is approximately unchanged when varying J.

In Experiment 2, we evaluate the ability of the algorithm to detect tandem repeats with gap allowed in multiple sequences named Short Tandem Repeats (STRs) markers. Also, we com-

pare our outcome with the result of testing those markers one after one using Tandem Repeats Finder (TRF) program [2].

6.2 Experiment on Synthetic Data

The settings of the synthetic DNA sequences are $N = 10$, $L = 5000$, $\Omega_{min} = 10$, $\Omega_{max} = 30$, $G = 3$, $J = 6$, and Θ written as follows,

$$\Theta = \begin{bmatrix} 0.80 & 0.10 & 0.10 & 0.05 & 0.10 & 0.85 \\ 0.05 & 0.80 & 0.15 & 0.05 & 0.80 & 0.05 \\ 0.10 & 0.05 & 0.70 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.85 & 0.05 & 0.05 \end{bmatrix}.$$

The settings of the algorithm are $G = 3$, $\varepsilon = 0.25$, $Q_1 = Q_2 = Q_3 = 1/6$, and $Q_4 = Q_5 = 1/3$.

Table 6.1 and Table 6.2 shows the comparison of actual and predicted values of repeat segment starting positions and repeat unit copy numbers, respectively. Table 6.3 shows the average false positive and false negative after 100 independent repeated trials.

Table 6.1: Comparison of actual and predicted values of repeat segment starting positions.

Seq. No.	1	2	3	4	5
Actual	3928	4367	613	4404	3049
Predicted	3928	4367	613	4398	3032
Seq. No.	6	7	8	9	10
Actual	471	1343	2637	4617	4652
Predicted	480	1335	2637	4655	4652

Table 6.2: Comparison of actual and predicted values of repeat unit copy numbers.

Seq. No.	1	2	3	4	5	6	7	8	9	10
Actual	15	11	15	30	10	20	23	23	24	29
Predicted	15	11	15	31	12	21	26	23	18	29

Table 6.3: Average false positive (FP) and false negative (FN).

# of Actual Repeat Units	200	FN	21.23
Average # of Predicted Repeat Units	208.57	FP	29.80

Figure 6.1, Figure 6.2 and Figure 6.3 show the trace of the unnormalized joint posterior probability, the traces of all repeat segment starting positions and the traces of all repeat unit copy numbers, respectively. As shown in Figure 6.1, the MCMC trace initially escalates as the number of iterations increases, but after about 1200 iterations it stabilizes around the actual value of the jointly posterior probability. Figure 6.2 shows that in the first 100 iterations, the traces change dramatically and they converge to each stable state afterwards. Figure 6.3 shows the repeat unit copy numbers start going up from 1 and fluctuate within only a small range after about 1000 iterations. It is observed that the traces of repeat segment starting positions converges most fast. We also find that the traces of the unnormalized joint posterior probability and repeat unit copy numbers have almost simultaneously rising trend, especially around 1000 iterations, as the copy number traces of R_9 and R_{10} step up, the trace of the unnormalized joint posterior probability has a subsequent uprush.

In summary, the goals of the algorithm in Chapter 2 are

achieved using this synthetic data test.

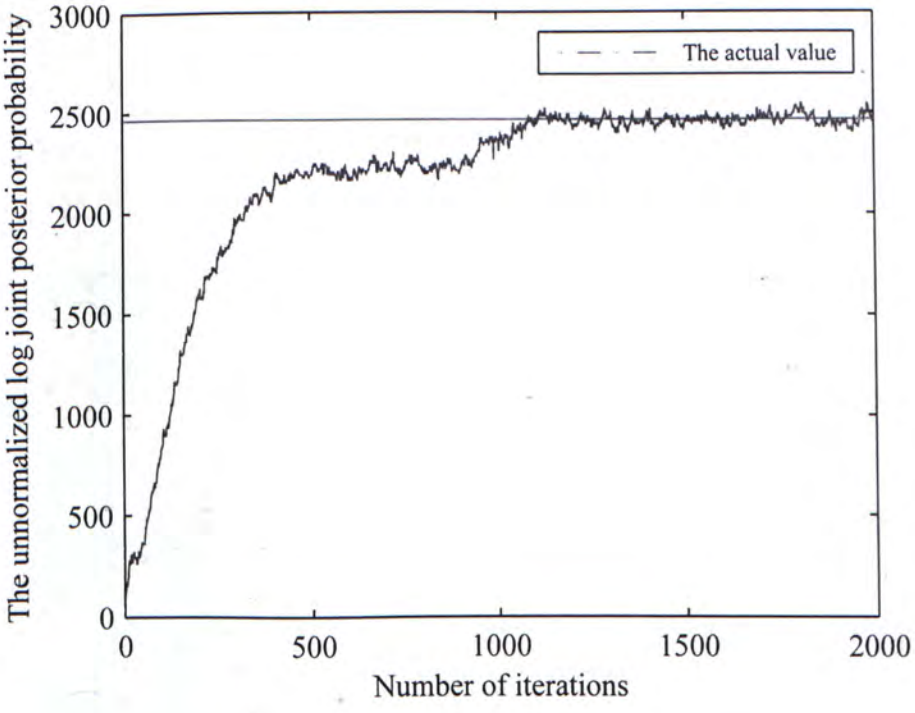

Figure 6.1: The trace plot of the unnormalized *log* joint posterior probability $P(\mathbf{A}, \mathbf{S}|\mathbf{R})$.

Since Θ and Φ are collapsed, the solution space of $\mathbf{A}$ and $\mathbf{S}$ is approximately unchanged when varying J. Based on that fact, we can execute the algorithm with a range of plausible widths and then choose the best result according to the MAP. We take a simple experiment that calculating each average unnormalized MAP after 100 independent repeated trials conditional on

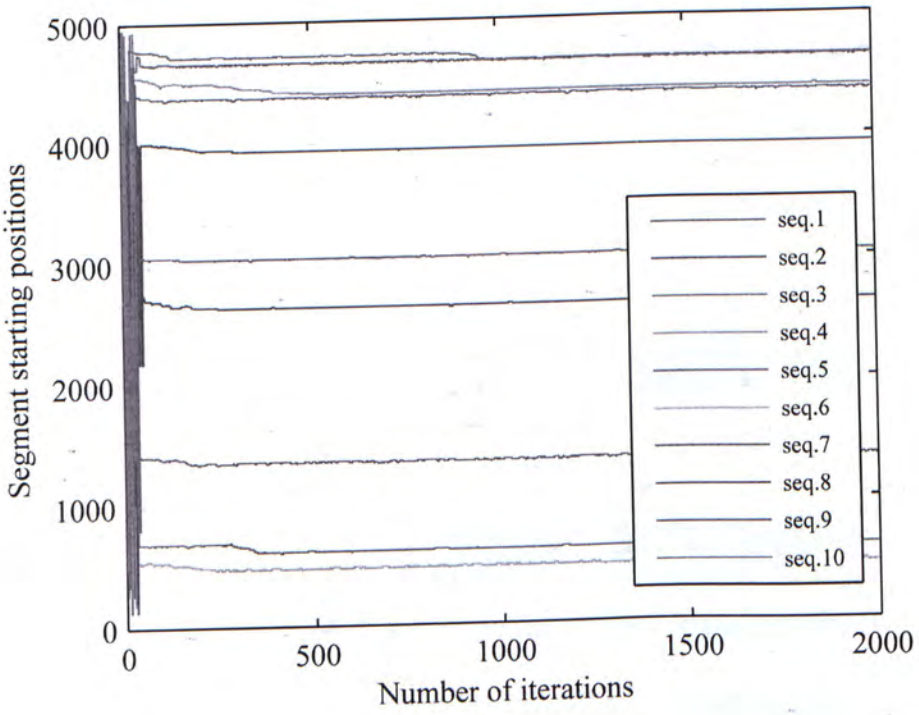

Figure 6.2: The trace plot of repeat segment starting positions $a_n, 1 \leq n \leq N$.

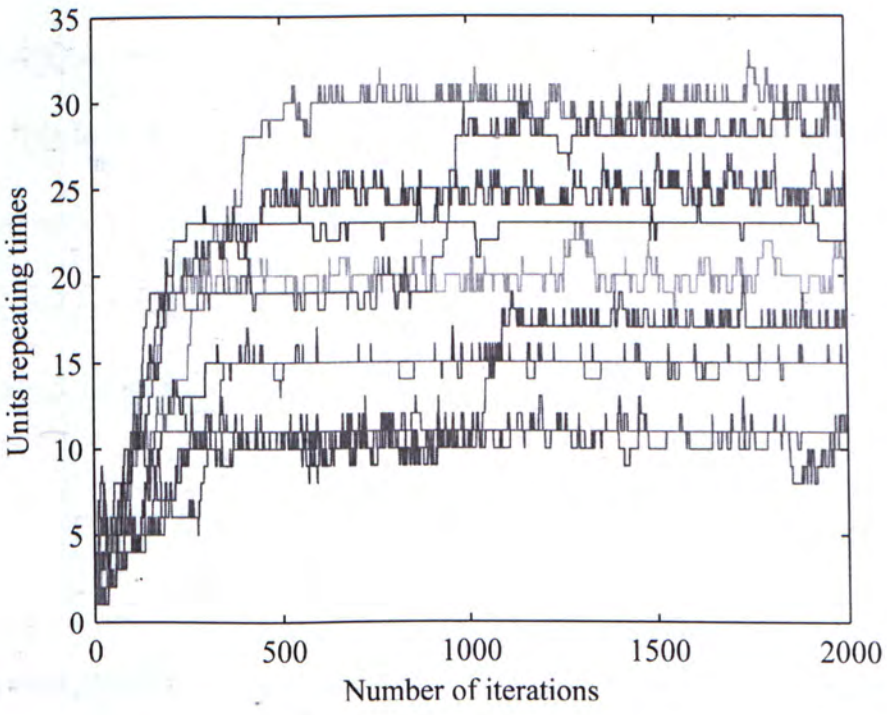

Figure 6.3: The trace plot of repeat unit copy numbers $|s_n|, 1 \leq n \leq N$.

different J and the result is shown in Figure 6.4. Clearly, the most likely value of J is 6. We also find that the other peaks approximately equal to 6-fold numbers.

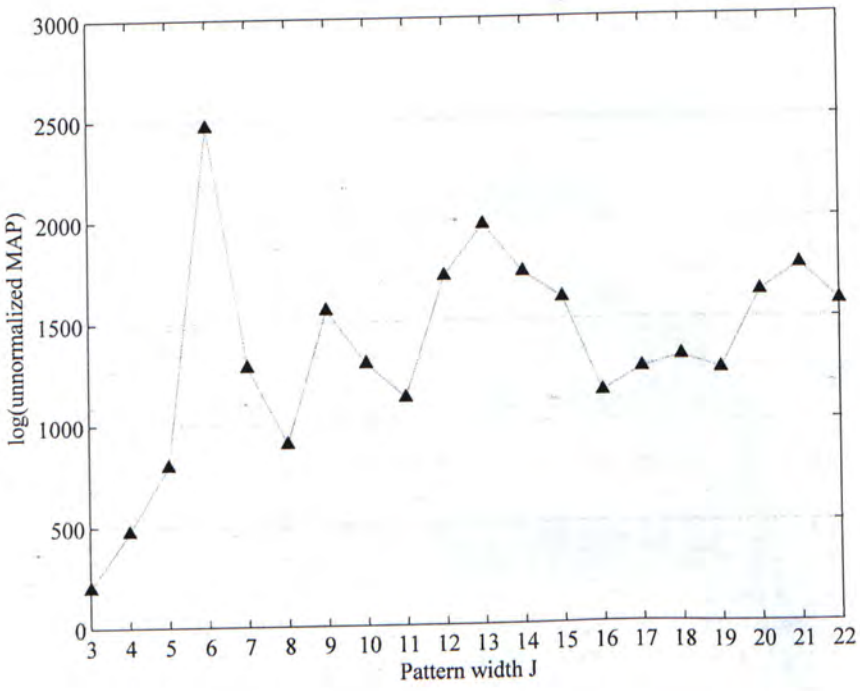

Figure 6.4: Average unnormalized *log* MAP conditional on different pattern width J.

6.3 Experiment on Real Data

We run our experiment on 6 Short Tandem Repeats (STRs) markers that have the length ranging from 131 bp to 420 bp and

the same pattern $GATA$ or its shifted form $ATAG$. With other 7 STR markers where the sequence patterns differ from $GATA$, this core set of 13 STR markers are being used to generate the FBI Combined DNA Index System (CODIS) database which has been successful at linking DNA profiles from crime scene evidence and at aiding paternity testing [31].

Table 6.4 presents the experiment result for these 6 STR markers via the algorithm with the setting $G = 3$, $\varepsilon = 0.10$, $Q_1 = Q_2 = Q_3 = 1/6$, and $Q_4 = Q_5 = 1/3$. The estimated motif matrix is as follows

$$\hat{\Theta} = \begin{bmatrix} 0.93 & 0.03 & 0.94 & 0.02 \\ 0.03 & 0.90 & 0.04 & 0.04 \\ 0.00 & 0.06 & 0.00 & 0.00 \\ 0.04 & 0.01 & 0.02 & 0.94 \end{bmatrix}.$$

For reference, TRF is used to detect tandem repeats within each marker and the result is also shown in Table 6.4. We find that the locations and the copy numbers of detected tandem repeats using TRF is a little bit different from those using our algorithm. But one of our superiorities is using a probabilistic matrix to

model tandem repeats rather than only reporting repeat unit pattern.

Table 6.4: Tandem repeats identified by TRF and our algorithm.

	TRF			Ours	
Marker	Pattern	Location	Copy No.	Location	Copy No.
CSF1PO	*AGAT*	94	18	92	16
D3S1358	*AGAT*	41	16.8	39	16
D5S818	*AGAT*	110	13.8	112	13
D7S820	*GATA*	126	15.3	125	15
D13S317	*GATA*	128	17.3	94	24
D16S539	*GATA*	275	11	261	16

□ **End of chapter.**

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we expand the scope of identifying tandem repeats to multiple DNA sequences, by relaxing the implicit assumption of a single DNA sequence in existing work, which are either based on string matching methods or signal processing methods. This is helpful in analyzing the relationship among DNA sequences, e.g., different species, in the course of evolution.

Also, we introduce a full probabilistic generative model to model the tandem repeats. As far as we know, this is the first work using sequence motif model Θ for the problem of detecting

repeats from a homogeneous background Φ. In order to handle gaps between repeating units, we design a binary vector data structure $\mathbf{s_n}$.

Furthermore, we introduce a Bayesian approach to detect tandem repeats in a de novo fashion. The MCMC algorithm is to iterate the sampling steps: $P(\Theta, \Phi|\mathbf{A}, \mathbf{S}, \mathbf{R})$, $P(\mathbf{A}|\mathbf{S}, \Theta, \Phi, \mathbf{R})$, and $P(\mathbf{S}|\mathbf{A}, \Theta, \Phi, \mathbf{R})$, each of which updates the corresponding parameter conditional on the current values of other parameters. After the MCMC iterations converge, the sampled parameter values give us a whole picture of their jointly posterior distribution. In order to improve computing efficiency, we use a collapsing technique by reducing the parameter space to only two kinds of parameters, $\mathbf{A}$ and $\mathbf{S}$. We demonstrate the effectiveness of the algorithm through experiments on both synthetic data and real data.

7.2 Future Work

We believe that Bayesian approach for identifying tandem repeats in multiple DNA sequences can find many applications. Our work takes the initial step to enable this repeats identification across multiple DNA sequences services. Many interesting and important directions are worth exploring. For example, our work is limited in the sense that requiring the repeat unit with the same width, in other words, not allowing the case of deletions and inter-unit insertions. Another question is how to implement parallel computing on the MCMC algorithm so as to make it converged into global optimum faster. Studying these problems is an interesting future direction.

□ **End of chapter.**

Bibliography

[1] J. Armour, T. Anttinen, C. May, E. Vega, A. Sajantila, J. Kidd, K. Kidd, J. Bertranpetit, S. Paabo, and A. Jeffreys. Minisatellite diversity supports a recent African origin for modern human. *Nature Genetics*, 13:154–160, 1996.

[2] G. Benson. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27(2), 1999.

[3] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973.

[4] V. Boeva, M. Regnier, D. Papatsenko, and V. Makeev. Short fuzzy tandem repeats in genomic sequence, identi-

fication and possible role in regulation of gene expression. *Bioinformatics*, 22(6):676–684, 2006.

[5] T. A. Brown. *Genomes*. New York: Garland Science, 2007.

[6] M. Buchner and S. Janjarasjitt. Detection and visualizaiton of tandem repeats in DNA sequences. *IEEE Transactions on Signal Processing*, 51(9):2280–2287, 2003.

[7] J. M. Butler, C. M. Ruitberg, and D. J. Reeder. STRBase: A short tandem repeat DNA internet-accessible database. In *Proceeding of the 8th International Symposium on Human Identification*, pages 38–47. Promega, 1997.

[8] V. Campuzano, L. Montermini, M. D. Molto, L. Pianese, and M. Cossee. Friedreichs ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, 271:1423–1427, 1996.

[9] A. T. Castelo, W. Martins, and G. R. Gao. TROLL-tandem repeat occurrence locator. *Bioinformatics*, 18:634–636, 2002.

[10] J. R. Collins, R. M. Stephens, B. Gold, B. Long, M. Dean, and S. K. Burt. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*, 82:10–19, 2003.

[11] J. Du, Y. Zhu, A. Shanmugam, and A. L. Kenter. Analysis of immunoglobulin SGAMMA3 recombination breakpoints by PCR: Implications for the mechanism of isotype switching. *Nucleic Acids Research*, 25:3066–3073, 1997.

[12] A. Edwards, H. Hammond, L. Jin, C. Caskey, and R. Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12:241–253, 1992.

[13] Y. H. Fu, A. Pizzuti, J. Fenwick, R. G. King, S. Rajnarayan, P. W. Dunne, J. Dubel, G. A. Nasser, T. Ashizawa, P. DeJong, B. Wieringa, R. Korneluk, M. B. Perryman, H. F. Epstein, and C. T. Caskey. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Science*, 255:1256–1258, 1992.

[14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* New York: Chapman & Hall / CRC, 2004.

[15] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[16] R. Gupta, D. Sarthi, A. Mittal, and K. Singh. A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequence. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:1–8, 2007.

[17] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–99, 1970.

[18] A. L. Hughes and H. Piontkivska. DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. *BMC Evolutionary Biology*, 5(12), 2005.

[19] L. Hunter. *Artificial Intelligence and Molecular Biology.* Cambridge: The MIT Press, 1993.

[20] Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983, 1993.

[21] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[22] B. Lewin. *Genes.* Oxford: Oxford University, 1997.

[23] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* New York: John Wiley, 1987.

[24] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* New York: Springer-Verlag, 2001.

[25] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs

sampling strategies. *Journal of the American Statistical Association*, 90(432):1156–1170, 1995.

[26] Q. Lu, L. L. Wallrath, H. Granok, and S. C. Elgin. $(CT)_n\,(GA)_n$ repeats and heat shock elements have distinct roles in chromation structure and transcriptional activation of the drosophila HSP26 gene. *Molecular and Cellular Biology*, 13:2802–2814, 1993.

[27] M. G. Main and R. J. Lorentz. An O(*nlogn*) algorithm for finding all repetitions in a string. *Journal of Algorithms*, 5(3):422–432, 1984.

[28] M. G. Main and R. J. Lorentz. Linear time recognition of square free strings. *Combinatorial Algorithms on Words, NATO ASI Series, Series F: Computer and System Sciences*, 12:272–278, 1985.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[30] E. Rivals, O. Delgrange, M. Dauchet, J. P. Delgrange, M. O. Delorme, A. Henaut, and E. Ollivier. Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequence. *Computer Applications in the Biosciences*, 13:131–136, 1997.

[31] C. M. Ruitberg, D. J. Reeder, and J. M. Butler. Strbase: A short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research*, 29(1):320–322, 2001.

[32] M. F. Sagot and E. W. Myers. Identifying satellites and periodic repetitions in biological sequences. *Journal of Computational Biology*, 5:539–553, 1998.

[33] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy. Spectral repeat finder (SRF): Identification of repetitive sequences using fourier transformation. *Bioinformatics*, 20(9):1405–1412, 2004.

[34] R. R. Sinden. Trinucleotide repeats biological implication of the DNA structures associated with disease-causing triplet

repeats. *Human Genetics*, 64:346–353, 2000.

[35] R. R. Sinden, V. N. Potaman, E. A. Oussatcheva, C. E. Pearson, Y. L. Lyubchenko, and L. S. Shlyakhtenko. Triplet repeat DNA structures and human genetic disease: Dynamic mutations fromdynamic DNA. *Journal of Biosciences*, 27(1):53–65, 2002.

[36] E. Y. Siyanova and S. M. Mirkin. Expansion of trinucleotide repeats. *Molecular Biology*, 35(2):168–182, 2001.

[37] A. L. Spada, E. Wilson, D. Lubahn, A. Harding, and K. Fischbeck. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, 352:77–79, 1991.

[38] T. Strachan and A. P. Read. *Human Molecular Genetics.* New York: John Wiley, 1999.

[39] D. Sussillo, A. Kundaje, and D. Anastassiou. Spectrogram analysis of genomes. *EURASIP Journal on Advances in Signal Processing*, 2004(1):29–42, 2004.

[40] G. R. Sutherland and R. I. Richards. Simple tandem DNA repeats and human genetic disease. *National Academy of Sciences*, 92(9):3636–3641, 1995.

[41] M. A. Tanner and W. H. Wong. The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association*, 80:528–550, 1987.

[42] S. A. Tishkoff, E. Dietzsch, W. Speed, A. J. Pakstics, and J. R. Kidd. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271:1380–1387, 1996.

[43] T. T. Tran, V. A. Emanuele, and G. T. Zhou. Techniques for detecting approximate tandem repeats in DNA. In *Proceeding of the IEEE International Conference on Acoustic Speech Signal Process*, volume 5, pages 449–452. IEEE, 2004.

[44] A. Verkerk, M. Pieretti, J. Sutcliffe, Y. Fu, D. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. Victoria, F. Zhang, B. Eussen, G. van Ommen, A. Blonden, G. Riggins, J. Chastain,

C. Kunst, H. Galjaard, C. Caskey, D. Nelson, B. Oostra, and S. Warren. Identification of a gene FMR1 containing a *CGG* repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile-X syndrome. *Cell*, 65:905–914, 1991.

[45] R. F. Weaver. *Molecular Biology*. Columbus: McGraw-Hill, 2004.

[46] J. Weber and P. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain-reaction. *American Journal of Human Genetics*, 44:388–396, 1989.

[47] M. Weitzmann, K. Woodford, and K. Usdin. DNA secondary structures and the evolution of hypervariable tandem arrays. *Journal of Biological Chemistry*, 272:9517–9523, 1997.

[48] R. Wells. Molecular basis of genetic instability of triplet repeats. *Journal of Biological Chemistry*, 271:2875–2878, 1996.

[49] L. Wu, J. Hong, and F. Lin. An approach to finding short tandem repeats in complete genomes. In *Proceeding of the German Conference on Bioinformatics 2001*, pages 238–241. GBF, 2001.

[50] H. X. Zhou, L. P. Du, and H. Yan. Detection of tandem repeats in DNA sequences based on parametric spectral estimation. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):747–755, 2009.